Bayesian Information Criterion for Singular Models

Mathias Drton

Department of Statistics
University of Chicago → Washington

Joint work with an AIM group:
Vishesh Karwa, Martyn Plummer, Aleksandra Slavkovic, Russ Steele
Outline

1. The Bayesian information criterion (BIC)
2. Singular models and ‘circular reasoning’
3. A proposal for a ‘singular BIC’
4. Examples
Outline

1. The Bayesian information criterion (BIC)
2. Singular models and ‘circular reasoning’
3. A proposal for a ‘singular BIC’
4. Examples
Bayesian information criterion (BIC)

- Observe a sample \( Y_1, \ldots, Y_n \)
- Parametric model \( \mathcal{M} \) (set of probability distributions \( \pi \))
- Maximized log-likelihood function \( \hat{\ell}(\mathcal{M}) \)

Bayesian information criterion (Schwarz, 1978)

\[
\text{BIC}(\mathcal{M}) := \hat{\ell}(\mathcal{M}) - \frac{\dim(\mathcal{M})}{2} \log n
\]

‘Generic’ model selection approach:

Maximize BIC(\( \mathcal{M} \)) over set of considered models
Example: Linear regression

- Observe $n$ independent realizations of the ‘system’:
  \[
  Y = \omega_1 X_1 + \omega_2 X_2 + \omega_3 X_3 + \epsilon \quad \text{with} \quad \epsilon \sim N(0, \sigma^2)
  \]

- Models $\equiv$ coordinate subspaces:
  \[
  M_J = \{ \omega \in \mathbb{R}^3 : \omega_j = 0 \ \forall j \notin J \}, \quad J \subseteq \{1, 2, 3\}.
  \]

- Dimension $|J|$ (or rather, $|J| + 1$)
Linear regression (covariates i.i.d. \( N(0, 1), \omega_1 = 1, \sigma = 2 \))
Motivation: 1) Bayesian model choice — Example

- \( Y_i = (Y_{i1}, Y_{i2}) \) : vector of two binary r.v., with joint distribution
  \[
  \pi = \begin{pmatrix}
  \pi_{11} & \pi_{12} \\
  \pi_{21} & \pi_{22}
  \end{pmatrix} \in \Delta_3.
  \]

- Models
  \[
  \mathcal{M}_1 = \{ \pi = \frac{1}{4} \cdot 1 \}, \quad \mathcal{M}_2 = \{ \pi \text{ of rank 1} \}, \quad \mathcal{M}_3 = \Delta_3.
  \]
Motivation: 1) Bayesian model choice — Example

- $Y_i = (Y_{i1}, Y_{i2})$: vector of two binary r.v., with joint distribution
  \[ \pi = \begin{pmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{pmatrix} \in \Delta_3. \]

- Models
  \[ M_1 = \{ \pi = \frac{1}{4} \cdot 1 \}, \quad M_2 = \{ \pi \text{ of rank 1} \}, \quad M_3 = \Delta_3. \]

- Prior on models, e.g.,
  \[ P(M_1) = P(M_2) = P(M_3) = \frac{1}{3}. \]
Motivation: 1) Bayesian model choice — Example

- \( Y_i = (Y_{i1}, Y_{i2}) \): vector of two binary r.v., with joint distribution
  \[
  \pi = \begin{pmatrix}
  \pi_{11} & \pi_{12} \\ 
  \pi_{21} & \pi_{22}
  \end{pmatrix} \in \Delta_3.
  \]

- Models
  \[
  M_1 = \{\pi = \frac{1}{4} \cdot 1\}, \quad M_2 = \{\pi \text{ of rank 1}\}, \quad M_3 = \Delta_3.
  \]

- Prior on models, e.g.,
  \[
  P(M_1) = P(M_2) = P(M_3) = \frac{1}{3}.
  \]

- Prior on \( \pi \) given model, e.g.,
  \[
  P(\pi | M_2) = \text{distribution on rank 1 matrices in } \Delta_3.
  \]
Motivation: 1) Bayesian model choice — Example

- \( Y_i = (Y_{i1}, Y_{i2}) \): vector of two binary r.v., with joint distribution

\[
\pi = \begin{pmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{pmatrix} \in \Delta_3.
\]

- Models

\( M_1 = \{ \pi = \frac{1}{4} \cdot 1 \} \), \( M_2 = \{ \pi \text{ of rank 1} \} \), \( M_3 = \Delta_3 \).

- Prior on models, e.g.,

\[
P(M_1) = P(M_2) = P(M_3) = \frac{1}{3}.
\]

- Prior on \( \pi \) given model, e.g.,

\[
P(\pi \mid M_2) = \text{distribution on rank 1 matrices in } \Delta_3.
\]

- Data-generating process under distribution \( \pi \) from \( M_i \):

\[
P(Y_1, \ldots, Y_n \mid \pi, M_i) = \prod_{i=1}^{n} \pi_{y_{i1}, y_{i2}}
\]
Motivation: 1) Bayesian model choice

- Posterior model probability in fully Bayesian treatment:

  \[ P(\mathcal{M} \mid Y_1, \ldots, Y_n) \propto P(\mathcal{M}) P(Y_1, \ldots, Y_n \mid \mathcal{M}). \]

- Marginal likelihood:

  \[ L_n(\mathcal{M}) := P(Y_1, \ldots, Y_n \mid \mathcal{M}) \]

  \[ = \int P(Y_1, \ldots, Y_n \mid \pi, \mathcal{M}) \, dP(\pi \mid \mathcal{M}) \]
Motivation: 2) Asymptotics

$Y_1, \ldots, Y_n$ i.i.d. sample from a distribution $\pi_0$ in a parametric model $M = \{\pi(\omega) : \omega \in \mathbb{R}^d\}$

**Theorem (Schwarz, 1978; Haughton, 1988; and others)**

Assume $P(\omega | M)$ is a ‘nice’ prior on $\mathbb{R}^d$. Then in ‘nice’ models,

$$\log L_n(M) = \hat{\ell}_n(M) - \frac{d}{2} \log(n) + O_p(1),$$

and a better (Laplace) approximation to error $O_p(n^{-1/2})$ is possible.
Motivation: 2) Asymptotics

\[ Y_1, \ldots, Y_n \text{ i.i.d. sample from a distribution } \pi_0 \text{ in a parametric model} \]

\[ \mathcal{M} = \{ \pi(\omega) : \omega \in \mathbb{R}^d \} \]

**Theorem (Schwarz, 1978; Haughton, 1988; and others)**

Assume \( P(\omega | \mathcal{M}) \) is a ‘nice’ prior on \( \mathbb{R}^d \). Then in ‘nice’ models,

\[
\log L_n(\mathcal{M}) = \hat{\ell}_n(\mathcal{M}) - \frac{d}{2} \log(n) + O_p(1),
\]

and a better (Laplace) approximation to error \( O_p(n^{-1/2}) \) is possible.

**Note:**

The ‘model complexity term’ \( \frac{d}{2} \log(n) \) does not depend on the true distribution \( \pi_0 \).
Where asymptotics come from: In a regular model . . .

- Gradient $\nabla \ell_n(\hat{\omega}) = 0$;
- Hessian $H_n(\hat{\omega})$ of $-\frac{1}{n}\ell_n$ converges to positive definite matrix (w.p. 1).
- Taylor approximation around MLE $\hat{\omega}$:
  \[
  \ell_n(\omega) \approx \ell_n(\hat{\omega}) - \frac{n}{2} (\omega - \hat{\omega})^\top H_n(\hat{\omega})(\omega - \hat{\omega})
  \]
- Integral approximately Gaussian (‘nice’ prior):
  \[
  \int_{\mathbb{R}^d} \exp(\ell_n(\omega)) \cdot f(\omega) \, d\omega \\
  \approx \exp(\ell_n(\hat{\omega})) \cdot f(\hat{\omega}) \cdot \int_{\mathbb{R}^d} \exp\left(-\frac{n}{2} (\omega - \hat{\omega})^\top H_n(\hat{\omega})(\omega - \hat{\omega})\right) \, d\omega \\
  \quad \cdot \sqrt{\frac{2\pi}{n}}^d \cdot \det(H_n(\hat{\omega}))^{-1}
  \]
Consistency

Theorem

Fix a finite set of ‘nice’ models. Then, BIC selects a true model of smallest dimension with probability tending to one as $n \to \infty$. 
Consistency

Theorem

Fix a finite set of ‘nice’ models. Then, BIC selects a true model of smallest dimension with probability tending to one as \( n \to \infty \).

Proof.

- Finite set of models \( \implies \) pairwise comparisons suffice.
- If \( P_0 \in \mathcal{M}_1 \subsetneq \mathcal{M}_2 \) and \( d_1 < d_2 \), then
  \[
  \hat{\ell}_n(\mathcal{M}_2) - \hat{\ell}_n(\mathcal{M}_1) = O_p(1); \quad \text{and} \quad (d_2 - d_1) \log n \to \infty.
  \]
- If \( P_0 \in \mathcal{M}_1 \setminus \text{clos}(\mathcal{M}_2) \), then with probability tending to one,
  \[
  \frac{1}{n} \left[ \hat{\ell}_n(\mathcal{M}_1) - \hat{\ell}_n(\mathcal{M}_2) \right] > \epsilon > 0; \quad \text{and} \quad \log(n)/n \to 0.
  \]
Outline

1 The Bayesian information criterion (BIC)
2 Singular models and ‘circular reasoning’
3 A proposal for a ‘singular BIC’
4 Examples
Singular models

- Model is singular if
  
  ‘approximation by a positive definite quadratic form not possible everywhere’.

- Examples:
  - Reduced rank regression
  - Gaussian mixture models
  - Any other mixture model
  - Factor analysis
  - Hidden Markov models
  - Graphical models with latent variables
  - …
Singular integrals

- In $\mathbb{R}^2$, ‘the’ regular integral behaves like
  \[
  \int_0^1 \int_0^1 e^{-n(x^2+y^2)} \, dx \, dy \sim \frac{\pi}{4} \cdot \frac{1}{n}
  \]

- An integral arising from a singular model could look like:
  \[
  \int_0^1 \int_0^1 e^{-nx^2y^2} \, dx \, dy \sim \frac{\sqrt{\pi}}{2} \cdot \frac{\log(n)}{n^{1/2}}.
  \]
Singular integrals

- In $\mathbb{R}^2$, ‘the’ regular integral behaves like

$$\int_0^1 \int_0^1 e^{-n(x^2+y^2)} \, dx \, dy \sim \frac{\pi}{4} \cdot \frac{1}{n}$$

- An integral arising from a singular model could look like:

$$\int_0^1 \int_0^1 e^{-nx^2y^2} \, dx \, dy \sim \sqrt{\frac{\pi}{2}} \cdot \frac{\log(n)}{n^{1/2}}.$$

Proof:

$$\int_0^1 \int_0^1 e^{-nx^2y^2} \, dx \, dy = \int_0^1 \frac{1}{y \sqrt{n}} \left( \int_0^{y \sqrt{n}} e^{-u^2} \, du \right) \, dy = \sqrt{\frac{\pi}{2}} \cdot \frac{1}{\sqrt{n}} \int_0^{\sqrt{n}} \frac{\text{erf}(v)}{v} \, dv$$

$$= \sqrt{\frac{\pi}{2}} \cdot \frac{1}{\sqrt{n}} \left[ \log(v)\text{erf}(v) \right]_0^{\sqrt{n}} - \frac{1}{\sqrt{n}} \int_0^{\sqrt{n}} \log(v)e^{-v^2} \, dv = \sqrt{\frac{\pi}{2}} \cdot \frac{\log(n)}{n^{1/2}} + O\left(\frac{1}{\sqrt{n}}\right).$$

Substitution $u = (y \sqrt{n}) \cdot x$, then $v = \sqrt{n} \cdot y$, then integration by parts.
Watanabe’s theorem

Theorem 6.7 in Watanabe (2009)

Suppose $Y_1, \ldots, Y_n$ are drawn i.i.d. from a distribution $\pi_0$ in a singular model $\mathcal{M}$. Then, under ‘suitable technical conditions’, the marginal likelihood sequence satisfies

$$\log L_n(\mathcal{M}) = \hat{\ell}_n(\mathcal{M}) - \lambda(\pi_0) \log(n) + \left[ m(\pi_0) - 1 \right] \log \log(n) + O_p(1).$$
Watanabe’s theorem

Theorem 6.7 in Watanabe (2009)

Suppose \( Y_1, \ldots, Y_n \) are drawn i.i.d. from a distribution \( \pi_0 \) in a singular model \( \mathcal{M} \). Then, under ‘suitable technical conditions’, the marginal likelihood sequence satisfies

\[
\log L_n(\mathcal{M}) = \hat{\ell}_n(\mathcal{M}) - \lambda(\pi_0) \log(n) + \left[ m(\pi_0) - 1 \right] \log \log(n) + O_p(1).
\]

Note:

The ‘model complexity term’ \( \lambda(\pi_0) \log(n) - \left[ m(\pi_0) - 1 \right] \log \log(n) \) generally depends on the unknown true distribution \( \pi_0 \).
How might we use mathematical knowledge...?

- Example: reduced rank regression
- Parameter space
  \[
  \{ \text{matrices of rank } \leq H \}
  \]
- Model selection problem: Determine appropriate rank \( H \)
- Singularities \( \pi_0 \) correspond to matrices of rank \( < H \) (‘null set’)
- Learning coefficient \( \lambda(\pi_0) \) and its order \( m(\pi_0) \) are functions of rank of \( \pi_0 \).

‘Circular reasoning’

To define a truly Bayesian information criterion for singular models overcome:

Rank unknown \( \leftrightarrow \) Model complexity unknown
1. The Bayesian information criterion (BIC)

2. Singular models and ‘circular reasoning’

3. A proposal for a ‘singular BIC’

4. Examples
Setup

- Finite set of competing models: \( \{M_i : i \in I\} \); closed under intersection
- Inclusion ordering: \( i \preceq j \) if \( M_i \subseteq M_j \)
- If the true data-generating distribution \( \pi_0 \) was known, then Watanabe’s theorem suggests replacing the marginal likelihood \( L(M_i) \) by

\[
L'_{\pi_0}(M_i) := P(Y_n | \hat{\pi}_i, M_i) \cdot n^{-\lambda_i(\pi_0)}(\log n)^{m_i(\pi_0)-1}.
\]

Our approach

Assign a probability distribution \( Q_i \) for \( \pi_0 \) and approximate \( L(M_i) \) as

\[
L'_{Q_i}(M_i) := \int_{M_i} L'_{\pi_0}(M_i) \, dQ_i(\pi_0).
\]

- How to choose \( Q_i \)?
Candidates for $Q_i$ — conditioning on $\mathcal{M}_i$

- Typically, it holds in an ‘almost surely’ sense that

$$\lambda(\pi_0) = \frac{\dim(\mathcal{M}_i)}{2}, \quad m(\pi_0) = 1.$$ 

- Hence, if $\pi_0$ follows the posterior distribution given model $\mathcal{M}_i$, that is,

$$Q_i(\pi_0) = P(\pi_0 \mid \mathcal{M}_i, Y_n)$$

then averaging wrto. $Q_i$ gives the standard BIC.

- This choice of $Q_i$ does not reflect uncertainty wrto. models; condition solely on $\mathcal{M}_i$. 
Candidates for $Q_i$ — conditioning on $M_i$ and submodels

- We suggest choosing the posterior distribution for $\pi_0$ given all submodels of $M_i$, that is,

$$Q_i = P(\pi_0 \mid \{ M : M \subseteq M_i \}, Y_n) = \frac{\sum_{j \preceq i} P(\pi_0 \mid M_j, Y_n)P(M_j \mid Y_n)}{\sum_{j \preceq i} P(M_j \mid Y_n)}.$$

- The distribution $Q_i$ puts positive mass on submodels.
  (e.g., positive prob to rank 1 matrices when considering rank 2 matrices)

- **BUT**: What about $P(M_j \mid Y_n)$? These are the posterior prob’s we want to approximate in the first place.
Recursive structure

Singular model selection problems typically have the following features:

- The smallest model is regular.
  - Reduced rank regression: Rank 0
  - Latent class models: Independence of discrete r.v.
  - Factor analysis: Independence of normal r.v.
  - ...

- Learning coefficients (order) are ‘almost surely’ constant along submodels. Approximation at generic points in $\mathcal{M}_j \subset \mathcal{M}_i$:

$$L'_{ij} := P(Y_n | \hat{\pi}_i, \mathcal{M}_i) \cdot n^{-\lambda_{ij}} (\log n)^{m_{ij} - 1} > 0.$$

- Reduced rank regression: matrices of lower rank
- Latent class models: matrices of lower rank
- Factor analysis: smaller $\#$ factors
- ...

- Recursive approach starting from the smallest model.
An equation system

- For our proposed choice of $Q_i$:

$$L'(M_i) := L'_{Q_i}(M_i) = \frac{1}{\sum_{j \leq i} P(M_j | Y_n)} \cdot \sum_{j \leq i} L'_{ij} P(M_j | Y_n)$$

$$= \frac{1}{\sum_{j \leq i} L(M_j)P(M_j)} \cdot \sum_{j \leq i} L'_{ij} L(M_j)P(M_j).$$

- Plugging-in approximations gives the equation system:

$$L'(M_i) = \frac{1}{\sum_{j \leq i} L'(M_j)P(M_j)} \cdot \sum_{j \leq i} L'_{ij} L'(M_j)P(M_j), \quad i \in I.$$
Triangular system

- Assume $P(M_i) \propto \text{const}$ in the sequel.
- Clearing denominators we obtain the triangular system:

$$L'(M_i)^2 + \left[ \left( \sum_{j \prec i} L'(M_j) \right) - L'_{ii} \right] \cdot L'(M_i) - \sum_{j \prec i} L'_{ij} L'(M_j) = 0, \quad i \in I.$$ 

For smallest model, the positive solution is

$$L'(M_i) = L'_{ii} > 0.$$

- Equation system has unique positive solution.

**Definition (Singular BIC)**

$$\text{sBIC}(M_i) = \log L'(M_i),$$

where $(L'(M_i) : i \in I)$ solves the above equation system.
Properties

- Reduces to ordinary BIC if models are regular
- Consistency
- Closer to behavior of Bayesian procedures
Outline

1. The Bayesian information criterion (BIC)
2. Singular models and ‘circular reasoning’
3. A proposal for a ‘singular BIC’
4. Examples
Reduced rank regression

Multivariate linear model for $Y \in \mathbb{R}^N$, with covariate $X \in \mathbb{R}^M$:

$$Y = CX + \epsilon, \quad \epsilon \sim \mathcal{N}(0, I_N).$$

**Definition**

Reduced rank regression model for rank $0 \leq H \leq \min\{M, N\}$:

$$Y \sim \mathcal{N}(CX, I_n \otimes I_N), \quad C \in \mathbb{R}^{N \times M} \text{ and } \text{rank}(C) \leq H.$$

Parameter space:

$$\mathbb{R}_{H}^{N \times M} := \{ C \in \mathbb{R}^{N \times M} : \text{rank}(C) \leq H \}.$$

$$\dim (\mathbb{R}_{H}^{N \times M}) = H(N + M - H).$$

$$\text{Sing} (\mathbb{R}_{H}^{N \times M}) = \mathbb{R}_{H-1}^{N \times M}.$$
Bayesian rank selection

Parametrization:

\[ g : \mathbb{R}^{N \times H} \times \mathbb{R}^{H \times M} \rightarrow \mathbb{R}_H^{N \times M} \]

\[ (B, A) \mapsto BA. \]

- Maximal rank of the Jacobian of \( g \):

\[ \dim (\mathbb{R}_H^{N \times M}) = H(N + M - H). \]

- Singularities of the map \( g \) (rank-deficient Jacobian):

\[ (B, A) \text{ with either } \text{rank}(B) < H \text{ or } \text{rank}(A) < H. \]

Posterior distribution for rank (with priors \( P(H) \) and \( \psi_H(B, A) \)):

\[ P(H \mid Y) = P(H) \int L(BA)\psi_H(B, A) \, d(B, A). \]

Learning coefficients \( \rightarrow \) Aoyagi & Watanabe (2005)
Singular BIC in action

\( \left( \frac{4}{3}, 1, \frac{2}{3}, 0, \ldots \right) \)

\( N = 7, \ M = 9, \ r = 3, \ n = 10 \)

- BIC
- sBIC

Estimated rank
Singular BIC in action

\( (5/4, 1, 3/4, 1/2, \ldots ) \)

\( N = 10, \ M = 15, \ r = 4, \ n = 15 \)
Singular BIC in action

$N = 30, \ M = 20, \ r = 4, \ n = 50$
Univariate Gaussian mixtures

True Rank = 2, N = 15

Estimated rank

BIC
sBIC
Factor analysis

$m = 10$, $k = 2$, $n = 25$

BIC
sBIC
More factor analysis

$m = 15, k = 4, n = 50$
What do you think?